【SLAA】

Field, A. (2009) Discovering Statistics Using SPSS(pp.273-282)

8.4 Assumptions and things that can go wrong

8.4.1 Assumptions

■ロジスティック回帰分析においても線形回帰分析と同様の前提が必要である

線形性 (Linearity)

■線形回帰分析の場合には独立変数と従属変数間が線形関係にあることが前提となっていたが、ロジスティック回帰分析の場合は独立変数と従属変数間に線形関係がないことが想定される

■ロジスティック回帰分析は従属変数を対数変換したロジット値を分析に用いるが、この従属変数のロジット値と独立変数間には線形関係があることが前提として必要である

■従属変数のロジット値と独立変数間の線形関係を把握する方法としては、これらの変数の交互作用項が有意であることを確認する方法がある (詳しくは 8.8.1 にて説明)

残差の独立性 (Independence of errors)

■線形回帰分析の場合と同様、回帰式の残差が独立している必要がある

■残差同士が独立していない場合は、変数の各ケースが関連しあっている可能性があるためである

■残差の独立性が満たされていない場合は過分散 (overdispersion; 8.4.4) が生じる場合がある

多重共線性 (multicollinearity)

■線形回帰分析の場合と同様、高い相関がある複数の独立変数を分析に用いてはいけない

■線形回帰分析の場合には許容度、VIF、固有値及び交差積行列によって確認することができた

■ロジスティック回帰分析の場合もほぼ同様の指標を用いる (詳しくは 8.8.1 にて説明)

その他→SPSS Tip 8.1, 8.4.2, & 8.4.3 で説明する

■変数内のカテゴリにサンプルが当てはまらないセルがある場合や、回帰式によって従属変数が完全に予測されてしまう場合には標準誤差が極端に大きくなるため、正確な分析ができなくなる場合がある

[SPSS Tip]

■多くの分析方法ではパラメータを算出するために iterative process (反復計算) を行う (e.g., 因子分析)

■反復計算においては、パラメータの予測を立てた上で計算を繰り返してより適切なパラメータとなるように計算を行う過程を繰り返す

■反復計算の過程の終了は、それ以上計算をしてもパラメータが改善しない場合と計算が規定の回数以上行われた場合になることで生じる

■SPSS においては「反復回数の制限に達しましたが､対数最尤値及びパラメータは収束しませんでした」と表示されるが、これは一定回数反復計算を行ってみても計算によって算出されるパラメータが収束していない (安定していない) ことを意味しているため、分析結果は当てにならないことを示す

■このような場合は計算の反復回数を増やすこと、もしくは計算の収束性を緩くする必要がある

8.4.2 Incomplete information from the predictors

■トマトの食生活及び喫煙の有無からがんの発生を予測する場合、トマト (食べる・食べない)、たばこ

(喫煙・禁煙)、がんの有無 (ある・なし) の全てのカテゴリに対して当てはまるサンプルが必要である
■SPSS でこのようなサンプル分布を確認するには、クロス集計表 (chapter 18) を確認するとよい
■その際、「期待度数」が 1 以上であり、かつ全ての期待度数の 20%以上の「期待度数」が 5 以上であることを確認することがロジスティック回帰分析の前提となる (カイ二乗→Fisher の検定にすべき)
■上記のことはカテゴリカル変数だけでなく連続変数にも該当し、欠損セルが生じている場合は回帰係数の標準誤差が極端に大きくなってしまう。より正確な分析を行うためにはできるだけ多くのサンプル数を収集するにこしたことはない

8.4.3 Complete separation
■独立変数によって従属変数が完全に予測されてしまう場合には分析結果が正確にならなくなる
■例として、セキュリティシステムに記録された体重からその人物を予測する場合を挙げる。予測されるのは、その人物が空き巣である (= 1) かそうでない (= 0) と仮定する
■まず家にある程度成長した子どもがいる場合、その家には子どもの友人がやってくるため比較的様々な体重のデータが収集される
■この場合のロジスティック曲線に従うと、40kg 以下のデータはほぼ 100%子どもの友人のものであり、85kg 以上になるとほぼ 100%のデータが空き巣のものであることが示唆される
■それに対して、40kg から 85kg のデータについてはその値によって空き巣かどうかの確率がどの程度かが算出される (e.g., 60kg だとほぼ 50%の確率である)
■その後、その子どもがひとり立ちして以降家にペットとして複数の猫を飼った場合を想定すると、その猫と仲間の猫のデータが収集されるようになる
■この場合のロジスティック曲線に従うと、15kg 以下のデータはすべてが猫のものであり、40kg 以上のデータは全てが空き巣のものであることが示唆される
■ただし、この場合に 15kg から 40kg のケースが空き巣である確率を計算することはできない
■このような場合は回帰係数の標準誤差が大きくなり、安定した結果は得られない

8.4.4 Overdispersion
■ロジスティック回帰式によって予測される分散よりも実際の分散が大きいことを過分散と呼び、正確な分析を妨げる要因となる
■過分散が生じる理由としては従属変数間の相関が高すぎる (i.e., 独立性の前提が崩れる) 場合と、サンプルによって各従属変数に当てはまる確率が異なっている場合が考えられる (e.g., 8.4.3 の例において空き巣が多い地区と空き巣が少ない地区のデータが混在している場合、体重を考慮する前の段階でサンプルによって空き巣であるかどうかの確率が異なっていると考えられる)
■過分散が生じると (二項分布で期待されるよりも残差のばらつきが大きくなるため) 回帰係数の標準誤差が小さくなり、その結果として信頼区間が狭くなる (回帰係数の大きさ自体には影響しない)
■回帰係数の信頼区間はその変数の有意性検定に使用されるため、信頼区間が小さくなることによって各回帰係数が不当に有意になりやすくなってしまう (タイプ I エラーが生じる)
■SPSS では回帰式の適合度指標としてカイ二乗検定が行われるが、そのカイ二乗値と自由度の比が 1 より大きい (特に 2 以上) だと問題があると考えられる

■過分散の影響については、標準誤差を「カイ二乗値」もしくは「カイ二乗値と自由度の比」で割った時の値が大きい方の値を用いることによってその影響を小さくすることができる

8.5 Binary logistic regression: an example that will make you feel eel

従属変数 (離散変数)：[Cured]　　治癒するかどうか (治癒する = 1, 治癒しない = 0)
独立変数 (離散変数)：[Intervention] 医療的介入の有無 (介入する = 1, 介入しない = 0)
独立変数 (連続変数)：[Duration]　　医療的介入をするまでの期間

8.5.1 The main analysis

■線形回帰分析と同様に、同じサンプルのデータが行になるように入力する
■[Cured] を従属変数に投入する
■[Intervention]、[Duration] 及び[Intervention*Duration] を共分散に投入する
→今回は各独立変数の主効果に加えて独立変数の交互作用を検討するためであり、不要であればそれぞれの独立変数のみを投入する形で良い

8.5.2 Method of regression

■今回は最尤法の計算方法に従った変数増加法による重回帰分析を行う
※通常であれば変数減少法の方がタイプⅡエラーを起こしにくい点で優れている

8.5.3 Categorical predictors

■独立変数の中でカテゴリカルな離散変数を設定する必要がある
■デフォルトであれば線形回帰分析におけるダミー変数と同様に扱われるが (0 をベースラインとする)、その他にも方法によってカテゴリカルな独立変数の扱い方を変えることも可能である (Table 10.6)
■デフォルトで分析を、どのカテゴリをベースラインとするかの設定が必要となる
※何をベースラインにするかによって解釈が異なるので、設定には注意が必要

8.5.4 Obtaining residuals

■線形回帰分析と同様にオプションで指定することにより残差を算出することができる
■ロジスティック回帰分析に特徴的なものとしては、predicted probabilities (予測された確率)、predicted group membership (予測されたグループ) がある
■前者は各ケースが従属変数の各カテゴリに当てはまる確率 (i.e., 治癒する確率)、後者は各ケースが従属変数の中で最も当てはまる可能性が高いカテゴリを示す
■モデルの適合率を調べることは統計分析において欠かせないことなので、必ず行うべきである

8.5.5 Further options

■SPSS のステップワイズにおける変数の減少及び増加の基準及びパラメータを収束させるための計算の反復回数が示されている
■これらの基準についてはデフォルトの状態で必要十分であるため、特別な理由がない限りは変更を行

わない方が良い

■回帰式に切片を含めたくない場合には、分析から除外させることも可能である

■Classification plot (分類プロット) にチェックを入れることで、ケースごとの実際の値と予測された値のヒストグラムを表示させることができる

■また、チェックを入れることで標準化残差が±2SD 以上であるケースを示すことができる

■ただし、SPSS で停止ざれる標準化残差が±2SD のケースのみを検討するだけでは十分とは言えないので、可能であれば残差の観点でデータを概観する方が望ましい

■さらに、各独立変数の 95%信頼区間及びモデルの適合度 (R^2) もチェックを入れることで出力できる

■上記のオプションに関しては基本的には出力することが望ましいと考えられる